Biomedical Optics EXPRESS

Classification of basal cell carcinoma in human skin using machine learning and quantitative features captured by polarization sensitive optical coherence tomography

TAHEREH MARVDASHTI,[1] LIAN DUAN,[1] SUMAIRA Z. AASI,[2] JEAN Y. TANG,[2] AND AUDREY K. ELLERBEE BOWDEN[1,*]

[1] *E. L. Ginzton Laboratory and Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA*
[2] *Department of Dermatology, Stanford University, Stanford, CA 94305, USA*
[*] *audrey@ee.stanford.edu*

Abstract: We report the first fully automated detection of basal cell carcinoma (BCC), the most commonly occurring type of skin cancer, in human skin using polarization-sensitive optical coherence tomography (PS-OCT). Our proposed automated procedure entails building a machine-learning based classifier by extracting image features from the two complementary image contrasts offered by PS-OCT, intensity and phase retardation (PR), and selecting a subset of features that yields a classifier with the highest accuracy. Our classifier achieved 95.4% sensitivity and specificity, validated by leave-one-patient-out cross validation (LOPOCV), in detecting BCC in human skin samples collected from 42 patients. Moreover, we show the superiority of our classifier over the best possible classifier based on features extracted from intensity-only data, which demonstrates the significance of PR data in detecting BCC.

OCIS codes: (110.4500) Optical coherence tomography; (170.3880) Medical and biological imaging; (230.5440) Polarization-selective devices; (100.0100) Image processing; (170.1870) Dermatology.

References and links

1. R. S. Stern, "Prevalence of a history of skin cancer in 2007: Results of an incidence-based model," Arch. Dermatol. **146**, 279–282 (2010).
2. UCSF School of Medicine, "Nonmelanoma skin cancer vs. melanoma," http://dermatology.medschool.ucsf.edu/skincancer/general/MelanomavNon.aspx.
3. M. A. Boone, S. Norrenberg, G. B. Jemec, and V. Del Marmol, "Imaging actinic keratosis by high-definition optical coherence tomography. Histomorphologic correlation: a pilot study," Exp. Dermatol. **22**, 93–97 (2013).
4. D. Cunha, T. Richardson, N. Sheth, G. Orchard, A. Coleman, and R. Mallipeddi, "Comparison of ex vivo optical coherence tomography with conventional frozen-section histology for visualizing basal cell carcinoma during Mohs micrographic surgery," Brit. J. Dermatol. **165**, 576–580 (2011).
5. D. Huang, E. Swanson, C. Lin, J. Schuman, W. Stinson, W. Chang, M. Hee, T. Flotte, K. Gregory, C. Puliafito, and a. et, "Optical coherence tomography," Science **254**, 1178–1181 (1991).
6. A. F. Fercher, "Optical coherence tomography," J. Biomed. Opt. **1**, 157–173 (1996).
7. R. Leitgeb, C. Hitzenberger, and A. F. Fercher, "Performance of fourier domain vs. time domain optical coherence tomography," Opt. Express **11**, 889–894 (2003).
8. T. Gambichler, A. Orlikov, R. Vasa, G. Moussa, K. Hoffmann, M. Stücker, P. Altmeyer, and F. G. Bechara, "In vivo optical coherence tomography of basal cell carcinoma," J. Dermatol. Sci. **45**, 167–173 (2007).
9. M. A. Boone, S. Norrenberg, G. B. Jemec, and V. Del Marmol, "Imaging of basal cell carcinoma by high-definition optical coherence tomography: histomorphological correlation. A pilot study," Brit. J. Dermatol. **167**, 856–864 (2012).
10. O. Markowitz, M. Schwartz, E. Feldman, A. Bienenfeld, A. K. Bieber, J. Ellis, U. Alapati, M. Lebwohl, and D. M. Siegel, "Evaluation of optical coherence tomography as a means of identifying earlier stage basal cell carcinomas while reducing the use of diagnostic biopsy," J. Clin. Aesthet. Dermatol. **8**, 14 (2015).
11. M. A. Boone, A. Marneffe, M. Suppa, M. Miyamoto, I. Alarcon, R. Hofmann-Wellenhof, J. Malvehy, G. Pellacani, and V. Del Marmol, "High-definition optical coherence tomography algorithm for the discrimination of actinic keratosis from normal skin and from squamous cell carcinoma," J. Eur. Acad. Dermatol. Venereol. **29(8)** 1–10 (2015).
12. T. M. Jorgensen, A. Tycho, M. Mogensen, P. Bjerring, and G. B. E. Jemec, "Machine-learning classification of non-melanoma skin cancers from image features obtained by optical coherence tomography," Skin Res. Technol. **14**,

#267313 http://dx.doi.org/10.1364/BOE.7.003721
Journal © 2016 Received 24 Jun 2016; revised 7 Aug 2016; accepted 8 Aug 2016; published 29 Aug 2016

364–369 (2008).

13. S. Schuh, R. Kaestle, E. C. Sattler, and J. Welzel, "Optical coherence tomography of actinic keratoses and basal cell carcinomas - differentiation by quantification of signal intensity and layer thickness," J. Eur. Acad. Dermatol. Venereol. (2016).
14. W. Gao, V. P. Zakharov, O. O. Myakinin, I. A. Bratchenko, D. N. Artemyev, and D. V. Kornilin, "Medical images classification for skin cancer using quantitative image features with optical coherence tomography," J. Innov. Opt. Health Sci. **9**, 1650003 (2016).
15. J. F. De Boer, T. E. Milner, M. J. van Gemert, and J. S. Nelson, "Two-dimensional birefringence imaging in biological tissue by polarization-sensitive optical coherence tomography," Opt. Lett. **22**, 934–936 (1997).
16. C. K. Hitzenberger, E. Götzinger, M. Sticker, M. Pircher, and A. F. Fercher, "Measurement and imaging of birefringence and optic axis orientation by phase resolved polarization sensitive optical coherence tomography," Opt. Express **9**, 780–790 (2001).
17. J. Strasswimmer, M. Pierce, B. Park, and V. Neel, "Polarization-sensitive optical coherence tomography of invasive basal cell carcinoma," J. Biomed. Opt. **9**, 292–298 (2004).
18. L. Duan, T. Marvdashti, A. Lee, J. Y. Tang, and A. K. Ellerbee, "Automated identification of basal cell carcinoma by polarization-sensitive optical coherence tomography," Biomed. Opt. Express **5**, 3717 (2014).
19. W. Trasischker, S. Zotter, T. Torzicky, B. Baumann, R. Haindl, M. Pircher, and C. K. Hitzenberger, "Single input state polarization sensitive swept source optical coherence tomography based on an all single mode fiber interferometer," Biomed. Opt. Express **5**, 2798–809 (2014).
20. K. L. Lurie, R. Angst, and A. K. Ellerbee, "Automated mosaicing of feature-poor optical coherence tomography volumes with an integrated white light imaging system," IEEE Trans. Biomed. Eng. **61**, 2141–2153 (2014).
21. C. A. Lingley-Papadopoulos, M. H. Loew, M. J. Manyak, and J. M. Zara, "Computer recognition of cancer in the urinary bladder using optical coherence tomography and texture analysis," J. Biomed. Opt. **13**, 024003–024003 (2008).
22. A. Miyazawa, M. Yamanari, S. Makita, M. Miura, K. Kawana, K. Iwaya, H. Goto, and Y. Yasuno, "Tissue discrimination in anterior eye using three optical parameters obtained by polarization sensitive optical coherence tomography," Opt. Express **17**, 17426–17440 (2009).
23. X. Qi, Y. Pan, M. V. Sivak, J. E. Willis, G. Isenberg, and A. M. Rollins, "Image analysis for classification of dysplasia in Barrett's esophagus using endoscopic optical coherence tomography," Biomed. Opt. Express **1**, 825–847 (2010).
24. Y. Yang, T. Wang, X. Wang, M. Sanders, M. Brewer, and Q. Zhu, "Quantitative analysis of estimated scattering coefficient and phase retardation for ovarian tissue characterization," Biomed. Opt. Express **3**, 1548–1556 (2012).
25. Y. Gan, D. Tsay, S. B. Amir, C. C. Marboe, and C. P. Hendon, "Automated classification of optical coherence tomography images of human atrial tissue," J. Biomed. Opt. **21**, 101407 (2016).
26. B. H. Park, C. Saxer, S. M. Srinivas, J. S. Nelson, and J. F. de Boer, "In vivo burn depth determination by high-speed fiber-based polarization sensitive optical coherence tomography," J. Biomed. Opt. **6**, 474–479 (2001).
27. P. Pande, S. Shrestha, J. Park, M. J. Serafino, I. Gimenez-Conti, J. Brandon, Y.-S. Cheng, B. E. Applegate, and J. a. Jo, "Automated classification of optical coherence tomography images for the diagnosis of oral malignancy in the hamster cheek pouch," J. Biomed. Opt. **19**, 086022 (2014).
28. K. W. Gossage, T. S. Tkaczyk, J. J. Rodriguez, and J. K. Barton, "Texture analysis of optical coherence tomography images: feasibility for tissue classification," J. Biomed. Opt. **8**, 570–575 (2003).
29. X. Qi, M. V. Sivak, G. Isenberg, J. E. Willis, and A. M. Rollins, "Computer-aided diagnosis of dysplasia in barrett's esophagus using endoscopic optical coherence tomography," J. Biomed. Opt. **11**, 044010 (2006).
30. P. B. Garcia-Allende, I. Amygdalos, H. Dhanapala, R. D. Goldin, G. B. Hanna, and D. S. Elson, "Morphological analysis of optical coherence tomography images for automated classification of gastrointestinal tissues," Biomed. Opt. Express **2**, 2821–2836 (2011).
31. J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," J. R. Stat. Soc. Ser. C Appl. Stat. **28**, 100–108 (1979).
32. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1226–1238 (2005).
33. J. Kittler, "Feature selection and extraction," in Handbook of Pattern Recognition and Image Processing pp. 59–83 (1986).
34. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of Ijcai **14**, 1137–1145 (1995).

1. Introduction

According to recent estimates, one in five Americans develop skin cancer in their lifetime [1]. Although not fatal, basal cell carcinoma (BCC) accounts for more than 80% of total incidences of skin cancer [2]. In the current workflow of the US medical system, visual assessment of suspicious skin lesions by a primary care physician (PCP) serves as the basis for referring a patient to a dermatologist for positive determination of BCC in biopsied skin. Among possible treatment options, Mohs micrographic surgery is the most effective and advanced, and involves consecutive removal with immediate evaluation of excised tissue via frozen section histopathology until clear margins are reached. Although proven successful, these diagnosis and treatment procedures suffer from several shortcomings. First, visual assessment of the suspicious lesion by PCP is a subjective and challenging task. Second, biopsy is an invasive procedure with a potentially disfiguring outcome. Moreover, histopathological evaluation of the biopsied skin is a time-consuming procedure. Given the availability of non-invasive BCC treatments such as photodynamic therapy [3] and pharmacological immunomodulation [3], invasive diagnostic techniques such as biopsy should be replaced by non-invasive alternatives. Lastly, Mohs surgery is a time-consuming and labor-intensive procedure, mainly due to the time needed for processing frozen-section histopathology [4]. Thus, there is a significant clinical need for a non-invasive, real-time, automated and reliable technique for BCC detection to assist the PCP to make accurate referrals and to replace the invasive biopsy and time-consuming histology process that contribute to lengthy Mohs operations.

Optical coherence tomography (OCT) is a non-invasive, label-free and non-contact imaging technique that captures volumetric images of subsurface structures of biological tissue [5–7]. The close-to-microscopy resolution (7 μm axially and 15 μm laterally), modest imaging depth (1-2 mm in tissue) and near real-time imaging that OCT offers enables measuring changes in skin layers including epidermis and dermis; hence, it is suitable for BCC detection. Previous publications have confirmed the efficiency of OCT to detect BCC in human skin through several *in-vivo* and *ex-vivo* clinical studies [8–11]. These publications have mostly suggested qualitative, rather than quantitative, features to identify BCC regions in OCT images; hence, the process is not automated. The work of Jorgensen et al. [12], is the closest attempt towards automating the BCC detection process; however, the use of visually extracted features from OCT images prevents full automation of their methodology. Moreover, their classifier achieved suboptimal performance (less than 80% specificity and 81% sensitivity), likely due to the low discriminative power of features that are extracted from normal OCT (intensity) images alone. Recently, Schuh et al. [13] attempted to detect BCC in OCT images through a quantitative procedure. However, their methodology requires manual selection of regions of interest and thus is not fully automated. Additionally, Gao et al. [14] recently investigated the discriminatory powers of several extracted quantitative features from OCT images in distinguishing BCC, melanoma and pigmented nevi. However, the accuracy of a final classifier was not reported.

The additional birefringence contrast of polarization-sensitive OCT (PS-OCT) [15, 16] can improve the specificity and sensitivity of detection of BCC, as the birefringence properties of normal skin are altered by the onset of BCC tumors [17]. Strasswimmer et al. [17] were the first to report the appearance of aggressive BCC in PS-OCT images and noted smaller values of phase retardation (PR) as a marker for regions showing aggressive BCC characteristics; however, since their preliminary work only included a small number of patients (two), the discriminative power of the proposed marker remains unclear. We recently reported a fully automated support vector machine-based classifier that identifies BCC in *ex-vivo* mice skin samples from PS-OCT images with high sensitivity and specificity (92.5% and 94.4%, respectively) [18]. The results confirmed that a classifier for BCC based on combined intensity and birefringence features outperforms a classifier based on intensity alone.

In this manuscript, we report a fully automated procedure to detect BCC in *ex-vivo* human skin (n = 42) from PS-OCT images. Using image processing techniques, we extracted numerous features from both intensity and polarization images. We then used machine learning to optimize the selection of features and to classify images as cancerous (BCC) or healthy using histopathology as the gold standard. The combination of intensity and birefringence features yielded a high sensitivity and specificity of 95.4% and 95.4%, respectively. To the best of our knowledge, the current study is the first fully automated diagnostic for BCC in human skin that uses PS-OCT images, and is the first demonstration of a large collection of PR image features to capture various birefringence properties of skin. The excellent accuracy of the classifier suggests the high discriminatory power of our proposed features, and the superior performance of a classifier including birefringence features suggests the relevance of PS-OCT data to skin cancer detection and possible automated detection of other tissue abnormalities that alter the birefringence properties of the tissue such as bladder cancer, breast cancer, burn depth, etc.

2. Materials and methods

2.1. Tissue sample collection

De-identified tissue samples were collected from the Stanford Dermatology Clinic in Redwood City, CA as part of the standard-of-care treatment for BCC (Mohs surgery). The study protocol was approved by Institutional Review Board (IRB) of Stanford University. The samples comprised unused pieces from the first stage of Mohs surgery operations confirmed (by prior biopsy) to contain BCC. For each patient, we collected one to four tissue samples. After excision, the tissue was frozen according to the standard-of-care frozen-section histology process. For some patients we also obtained surrounding tumor-free tissue ("dog-ears"), excised after the tumor is completely removed to facilitate in the reconstruction, which served as part of our healthy controls. Tissue samples were transported to the imaging facility on dry ice, defrosted and imaged by PS-OCT. All were imaged within six hours of excision and one hour of complete defrost.

We collected samples from 61 patients; however, samples from 19 patients were excluded based on the following exclusion criteria: 1) samples were heavily damaged due to unavoidable frozen-section histology incisions (n=7), 2) we were unable to confirm the histology match to a PS-OCT image (n=4), 3) PS-OCT images were heavily corrupted by image artifacts (e.g., photodetector saturation, n=7), or 4) the skin cancer was determined to be squamous cell carcinoma (SCC) and not BCC (n=1). Thus in this study, we included tissue samples from 42 patients.

2.2. PS-OCT imaging

The imaging system comprised a single-mode fiber (SMF), swept-source PS-OCT system based largely on a previous design [19]. The PS-OCT engine comprised a 20-kHz swept rate laser (Santec, HSL-2100) centered at 1325 nm. The system has an axial and lateral resolution of 8.9 μm and 12 μm, respectively, and a sensitivity of 99 dB. One or multiple (up to nine) volumes were acquired per sample to cover a majority of the sample. The volumetric data comprised $512 \times 256 \times 3648$ pixels in the X, Y and Z directions, respectively, and had a physical dimension of either 2.1×2.1 or 4.2×4.2 mm^2 in the XY-plane, and an imaging depth of 1.75 mm in Z.

2.3. Histology

Following imaging, we preserved the samples in 10% formalin for at least 72 hours. Afterwards, samples were shipped to a histopathology center for standard histology processing. Using ink marks as guidelines, each tissue sample was cut into 10-μm-thick sections, mounted on slides and stained with hematoxylin-eosin (H&E). A dermatologist (SZA) with histopathology expertise evaluated each slide to identify the presence of BCC.

Histology and PS-OCT images were matched by observing tissue ink marks visible in both images. Cuts and variations in tissue thickness served as additional visual clues for the matching process. An image was *labeled* BCC if the corresponding histology image contained a BCC tumor and *labeled* healthy otherwise. Histology was not acquired for dog-ear pieces, which we labeled as healthy, as they had been confirmed to be tumor-free. Based on histology results, our dataset comprised micro-nodular, nodular, superficial, and infiltrative subtypes of BCC.

2.4. Dataset construction

All processing and analysis was implemented in MATLAB R2014b. Reflectance, R, and single-pass PR, η, were calculated at each depth position z using standard PS-OCT equations [15]. We averaged adjacent complex Bscan data in sets of two to restrain noise before calculating reflectance and PR. We implemented algorithms involving intensity thresholding to find and exclude Ascans that were corrupted by image artifacts (e.g., photodetector saturation, deep cuts from the frozen-section histology process).

Table 1 summarizes the final dataset after post-processing and adjustments. Note that 10 patients contributed both healthy and BCC data, while for a different group of 16 patients we only obtained healthy data, and for another group of 16 patients we only obtained BCC data. Thus in total we included data from 42 patients. Additionally, depending on the physical sample size, the number of collected PS-OCT images varied significantly among patients. However, to avoid biasing our dataset towards patients with a larger number of collected PS-OCT images, we only retained ten PS-OCT images per type (healthy or cancerous), per patient. The smallest number of labeled images for any patient was ten, which set the upper bound for all patients. For patients yielding more than ten labeled images, ten images were picked randomly to keep in the dataset.

Table 1. Number of PS-OCT images (B-scans) and the total number of patients per class.* Not a miscalculation.

	# PS-OCT images	# patients
Healthy	260	26
BCC	260	26
Total	520	42*

Figure 1 shows examples of BCC and healthy PS-OCT images, with the corresponding histology. A healthy PR image (Fig. 1(b)) shows an increase in the PR values as a function of depth. Based on our observations, the rate of this increase can vary by patient and location on the body. Despite this variation, most PR images collected from healthy skin show roughly uniform PR patterns in the lateral direction. BCCs in human skin manifest diverse patterns in PR images. For example, a uniform pattern (similar to healthy skin) is observed in some subtypes of BCC, such as nodular BCC (Fig. 1(e)), whereas infiltrative BCC exhibits a different unique pattern (Fig. 1(h)).

2.5. Selecting the ROI

All features were calculated on data in a region of interest (ROI), defined in one of two ways. In both cases, the top boundary was determined by extracting the surface using a greedy algorithm [20] (Fig. 2(a)); the bottom boundary was selected to correspond to either a fixed distance (200 pixels = 940 μm) from the top surface (fixed ROI, Fig. 2(b)) or the deepest pixel attaining an intensity above a fixed threshold (binary ROI, Fig. 2(c)). The value of 200 pixels in the fixed ROI was chosen heuristically to greatly exclude pixels with low SNR. All Ascans within the fixed ROI have the same number of pixels; thus, this type of ROI is more appropriate for extracting features based on Ascans. The binary ROI comprises a binary image constructed from the intensity image by excluding values below a calculated threshold [21]. The binary ROI has the advantage of greatly restraining the noise and unreliable pixels (i.e., with lower SNR). For a given Bscan, each Ascan within the binary ROI may contain a different number of pixels.

2.6. Extracting features

Using both intensity and PR data, we extracted numerous image features from PS-OCT images. Our general strategy was to first extract multiple features to capture many properties of the skin and to then search for the subset of features that provides the highest accuracy when used to build

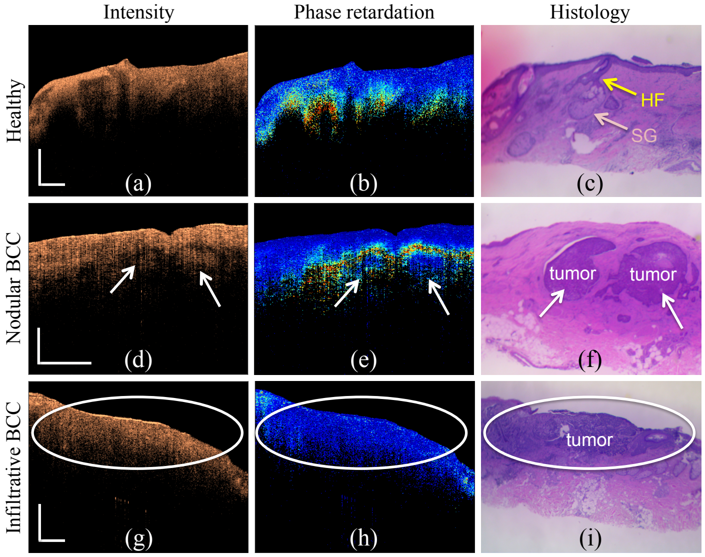

Fig. 1. Representative intensity (left), PR (middle) and histology (right) images of healthy and BCC human skin. (a), (b) and (c) represent the images of a healthy skin. Normal skin appendages such as a hair follicle (HF, yellow arrow) and sebaceous glands (SG, light pink arrow) are noticeable. (d), (e), and (f) images represent the case of a nodular BCC, and (g), (h) and (i) images represent the case of infiltrative BCC. The white arrows point to nodular tumor islands. The scale bars represent 500 µm × 500 µm and are applicable to all images in a row.

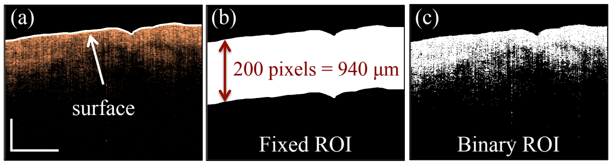

Fig. 2. Examples of (a) the extracted surface (serves as the top boundary) and the two types of ROI: (b) fixed ROI: 200 pixels (940 µm) beneath the surface, and (c) binary ROI: thresholding the intensity values (threshold value is calculated to be 100 dB here). The scale bars represent 500 µm × 500 µm.

a classifier. The features that we extracted can be divided into two main categories: 1) Features extracted from analyzing Ascans within a given Bscan, and 2) Features extracted from analyzing the whole Bscan. Table 2 summarizes the number of extracted features per category and the type of ROI used.

2.6.1. Ascan-based features

All Ascan-based features were calculated from the fixed ROI images. For all features, we first filtered both intensity and PR images using two-dimensional median filters with kernel sizes of [5 40] pixels (lateral vs. axial) and [25 40] for intensity and PR data, respectively. These filters sizes were chosen heuristically to simultaneously smooth the images and reinforce the layered structure of skin. Note that the fixed ROI was calculated from the original image and then applied to the filtered image. Examples of spatially filtered intensity and PR Bscans are presented in Figs. 3(a) and 3(c), respectively.

Table 2. Number of extracted features and type of ROI used for each feature category. Int: intensity, PR: phase retardation, F: fixed, and B: binary.

Ascan-based features				**Bscan-based features**			
	Int	**PR**	**ROI**		**Int**	**PR**	**ROI**
a. Long-range	12	28	F	a. Basic statistics	5	5	B
b. Short-range	120	124	F	b. Histogram	12	12	B
c. Peaks&valleys	40	44	F	c. Textures	48	96	B
d. Segment	36	52	F	d. Morphological	260	260	B
e. Crossing	76	76	F	e. Spatial freq.	4	4	F
Subtotal	284	324		Subtotal	329	377	
Total	608			Total	706		

Given the size of our dataset (i.e., 520 Bscan pairs) calculating Ascan-based features for each individual Ascan was deemed too time-consuming. Instead, for a given spatially filtered Bscan, we calculated Ascan features from laterally down-sampled images, comprising every fifth Ascan (since the Bscans were spatially averaged over five pixels for intensity data, and intensity filter has the smaller kernel compare to the kernel for PR data). Figure 3(b) and (d) present examples of down-sampled intensity and PR images, respectively.

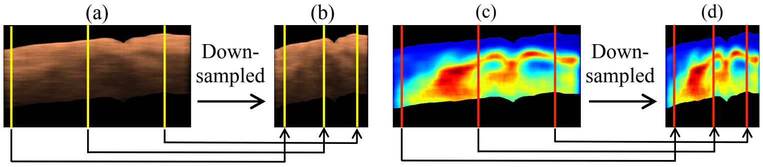

Fig. 3. Examples of filtered (a, and c) and down-sampled (b, and d) intensity and PR images, respectively. Down-sampled images are constructed by retaining every five Ascans (yellow or red colored Ascans).

Figure 4 illustrates the procedure for extracting the Ascan-based features. For each Ascan in the down-sampled image, we extracted *intermediate features* organized into an intermediate feature matrix (Fig. 4(b)): Columns and rows of the intermediate feature matrix represent Ascans and intermediate feature categories, respectively. For each intermediate feature category, we calculated four statistics (mean, standard deviation, minimum, and maximum) among all Ascans in a given Bscan (Fig. 4(c)); the values of these statistics themselves are the final Ascan-based features, organized into a one-dimensional vector (Fig. 4(d)). The same procedure was applied to PR images.

a. Long-range intermediate features To measure the attenuation coefficient [22–25], we used linear regression to fit a line to each intensity Ascan. We chose the pixel associated with the first peak of the intensity data as the starting pixel for the linear regression to exclude the stratum corneum and only capture epidermis properties. The slope, intercept and fitting error for this line were extracted as intermediate features. Previous publications have demonstrated the use of the slope of the line fitted to the PR Ascan (over a selected ROI) to detect tissue abnormalities and diseases such as basal cell carcinoma [17] and ovarian cancer, and to assess burn depth [26]. However, we chose to fit a higher order polynomial to the PR Ascan based on our observation that a higher order polynomial would provide a better fit. To exclude the stratum corneum, we chose the pixel associated with the first minimum of the PR data as the starting pixel for the polynomial fit. Then we fitted a 5th-order polynomial to these Ascans and extracted the six coefficients of the polynomial and the fitting error as intermediate features. The dashed red curves of Figs. 5(b) and

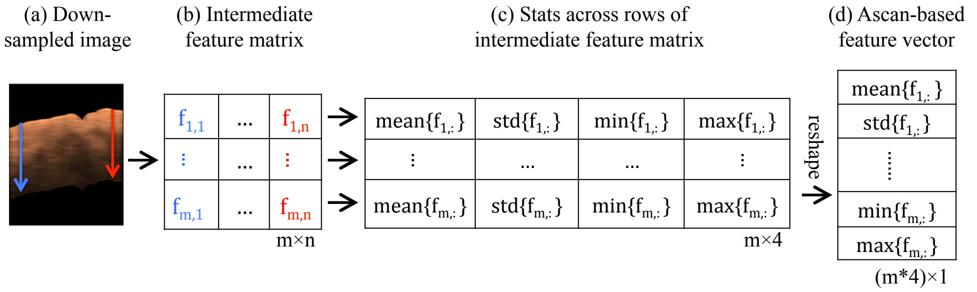
(a) Down-sampled image
(b) Intermediate feature matrix
(c) Stats across rows of intermediate feature matrix
(d) Ascan-based feature vector
$f_{1,1}$ ... $f_{1,n}$
⋮ ... ⋮
$f_{m,1}$ ... $f_{m,n}$
m×n
mean{$f_{1,:}$} std{$f_{1,:}$} min{$f_{1,:}$} max{$f_{1,:}$}
⋮ ... ... ⋮
mean{$f_{m,:}$} std{$f_{m,:}$} min{$f_{m,:}$} max{$f_{m,:}$}
m×4
reshape
mean{$f_{1,:}$}
std{$f_{1,:}$}
⋮
min{$f_{m,:}$}
max{$f_{m,:}$}
(m*4)×1

Fig. 4. Schematic illustration of the process for extracting Ascan-based features from a (a) down-sampled image: (b) construct the intermediate feature matrix by extracting several intermediate features from all Ascans (m=152), (c) calculate statistics (mean, std, min, and max) across the rows of intermediate feature matrix, and (d) construct final Ascan-based feature vector by reshaping the intermediate feature matrix into a vector.

5(h) illustrate examples of a line fitted to the intensity Ascan and a 5th-order polynomial fitted to the PR Ascan, respectively. In both figures the dashed black vertical lines represent the starting pixel for the fit.

b. Short-range intermediate features To extract local attenuation coefficient and birefringence properties, we performed a linear fit along a sliding axial window, 30 pixels long, along each intensity and PR Ascan. The shaded red rectangles and dashed lines of Figs. 5(c) and 5(i) represent the first sliding windows and the line fitted to the windowed data, respectively. Using linear regression, we extracted the slope, intercept and fitting error over the windowed values. We then calculated six statistics over all these slopes and intercept values as intermediate features: 1) mean, 2) median, 3) standard deviation, 4) mode, 5) minimum, and 6) maximum. Additionally, six intermediate features were calculated from the slope values: 1) difference between the slopes of the first and last windows [18], 2) slope of the best fitted line, 3-6) 0.25, 0.75, 0.90 and 0.95 quantiles. Moreover, we constructed histograms of these slope values and calculated histogram properties. The following twelve intermediate features were calculated from a given histogram: 1) skewness, 2) kurtosis, 3) median, 4) second moment, 5) third moment, 6) fourth moment, 7) standard deviation, 8) relative smoothness [21], 9) uniformity [21],, 10) entropy [21],, 11) height of the largest bin, and 12) the mean value of the bin boundaries associated with the largest bin. For the PR Ascan only, we also calculated the difference between the mean values of the first and second 50 pixels [18].

c. Peaks and valleys intermediate features Pande et al. observed loss of normal tissue layers as a hallmark for oral malignancy in the hamster cheek pouch [27]. To capture the number of tissue layers (i.e., degree of layering), they proposed features called “peaks and valleys” and “crossings.” For a given Ascan, these features provide a measure of the number of prominent peaks in the data. Here we extended their analysis in two ways: 1) by extracting similar features from PR data, and 2) by incorporating information about the axial positions of the peaks and valleys (with respect to the surface) as additional features. The procedure to localize peaks and valleys resembled that of Pande et al. [27], and involved constructing an axially flattened Ascan (black curves in Figs. 5(d) and 5(j)). Note that the axially flattened intensity Ascan was normalized. In Figs. 5(d) and 5(j), the peaks and valleys are marked by blue and green lines, respectively. We denoted the values of the *i*-th peak and *j*-th valley by P_i and V_j , respectively. We then located the *i*-th *local maximum* ($l_{max,i}$, blue upward-pointing triangle) and *j*-th *local minimum* ($l_{min,j}$, green downward-pointing triangle) as illustrated in Figs. 5(e) and 5(k) for intensity and PR Ascans, respectively, and denoted their axial positions by $z_{max,i}$ and $z_{min,j}$. The following parameters were calculated as intermediate features for both intensity and PR Ascans: 1)$\sum P_i$ [27], 2)$\sum P_i - \sum V_i$ [27], 3)$\sum P_i + \sum V_i$ [27], 4) number of located peaks (i.e., $\#P_i$), 5) number of located valleys (i.e., $\#V_i$), 6) standard deviation of $z_{max,i}$, 7) standard deviation of $z_{min,j}$, 8) value of the first local maximum ($l_{max,1}$), 9) value of the second local maximum ($l_{max,2}$), and 10) value of the first local minimum (i.e., $l_{min,1}$). Additionally, for the PR Ascan only we extracted the value of the second local minimum ($l_{min,2}$) as the last feature.

d. Segment intermediate features We defined *segments* as the portions of the Ascan bounded between each pair of consecutive local minima-local maxima. Upon visual observation, we recognized that segment patterns such as slopes, lengths, or total number of segments for a given Ascan, correlated with the state of the tissue (i.e., healthy vs. BCC). This correlation was especially strong for PR segments. Thus we extracted the segment patterns as another set of intermediate features. To some extent, the segment patterns capture local tissue properties, assuming a segment corresponds to a tissue layer or an axial change in tissue properties. Thus using linear regression, we fitted a line to each segment, as illustrated by the dashed red lines in Figs. 5(e) and 5(k). Denoting the slope of the fitted line to the *i*-th segment by $S_{g,i}$, the number of pixels within the i-th segment by $S_{l,i}$ (i.e., length of the *i*-th segment), and the total number of segments by N, we extracted the following intermediate features: 1) number of segments with positive slopes (i.e., $\#S_{g,i}$ for $S_{g,i} \geq 0$) , 2) number of segments with negative slopes (i.e., $\#S_{g,i}$ for $S_{g,i} < 0$), 3) mean of slope values for segments with positive slope values (i.e., mean of $S_{g,i}$ for $S_{g,i} \geq 0$) , 4) mean of slope values for segments with negative slope values (i.e., mean of $S_{g,i}$ for $S_{g,i} < 0$) , 5) $S_{g,1}$, 6) $S_{g,2}$, 7) $S_{g,3}$, 8) $S_{l,1}$, and 9) $S_{l,2}$. Four additional intermediate features were extracted for PR Ascans only: 10) weighted average of slope values defined as $\frac{1}{N} \sum_i S_{g,i} S_{l,i}$, 11) $S_{g,4}$, 12) $S_{l,3}$, and 13) $S_{l,4}$. Note that each Ascan could comprise a different number of segments. For consistency, we decided to only capture the slopes and lengths associated with the uppermost segments, as they represent the properties of the important skin layers for BCC detection such as epidermis. Additionally, our observation suggested that in general the PR Ascan comprised more segments than the intensity Ascan; therefore, we included more PR segment properties as intermediate features.

e. Crossing intermediate features We followed the same procedure as Pande et al. [27] for defining crossing intermediate features, which also aim at capturing the degree of tissue layering. The flattened Ascans were demarcated by 15 levels [27], illustrated by horizontal dashed lines in Figs. 5(f) and 5(l). We then defined C_i to be the number of times the flattened Ascan intersects level i. The following 19 parameters were defined as intermediate features: 1-15) C_i for $i = 1...15$, 16) mean of C_i , 17) median of C_i , 18) standard deviation of C_i , and 19) mode of C_i.

2.6.2. Bscan-based features

Bscan-based features, such as basic statistics [21], histogram properties [21], texture features [28, 29], morphological features [30], and spatial-frequency based features [28] have successfully been used to detect tissue abnormalities or diseases such as dysplasia in Barrett's esophagus [29], cancer in the urinary bladder [21], and oral malignancy [27] .

a. Basic statistics The following basic statistics [21] were used to define the first set of Bscan-based features for both intensity and PR data: 1) range, 2) standard deviation, 3) mean, 4) median, and 5) mode. These features were calculated only from values within the binary ROI.

b. Histogram statistics This set of features was calculated using the histogram of the intensity and PR data within the binary ROI. The intensity and PR histograms were constructed by

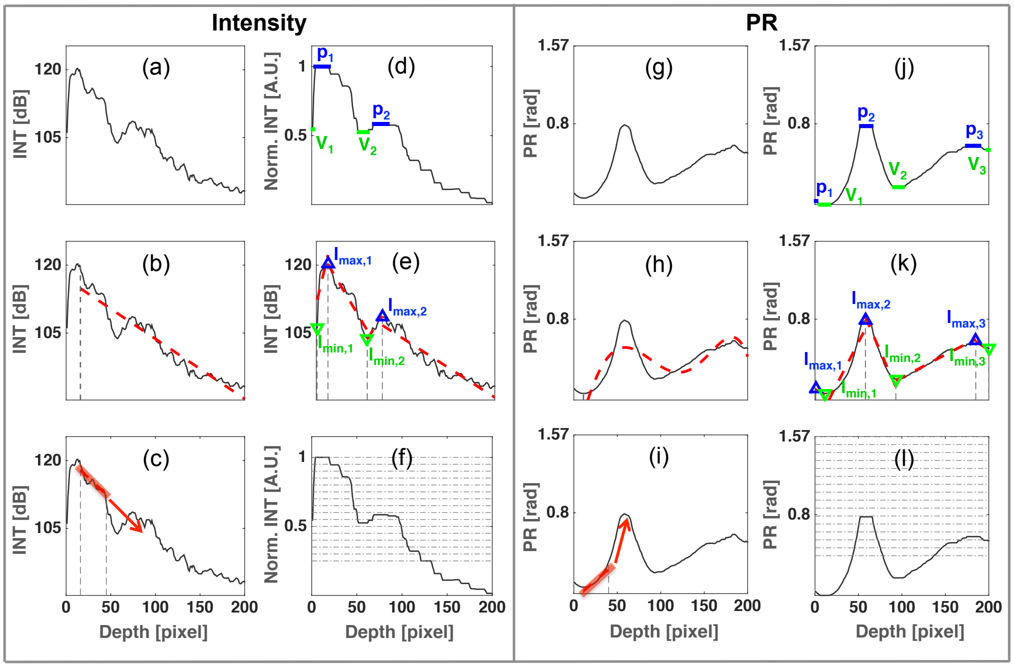

Fig. 5. Process for calculating all intermediate Ascan-based features from representative (a) intensity and (g) PR Ascans. In all images, dashed-red lines represent fitted lines, except in (h) where the dashed-red curve represents a fitted 5th-order polynomial. Vertical dashed-black lines represent the fitting ranges. Long-range intermediate features are calculated by fitting (b) a line to intensity and (h) a fifth-order polynomial to PR Ascans. Short-range intermediate features are calculated by fitting lines to data selected by an axially-moving window. The red rectangles represent the first window , and the window is translated axially (red arrows) in one-pixel steps. Peaks and valleys intermediate features are calculated from axially flattened (d) intensity and (j) PR Ascans. Blue and green lines represent peaks and valleys, respectively. Segment intermediate features are calculated by fitting lines to (e) intensity and (k) PR Ascans segments defined by local maxima (blue triangles) and local minima (inverted green triangles). Crossing intermediate features are calculated by counting the number of times (f) intensity and (l) PR Ascans cross predefined crossing levels (dashed horizontal lines). INT: intensity and Norm.: normalized.

binning the corresponding values into 64 and 16 bins, respectively. The twelve histogram features described in section 2.6.1(b) were extracted for both intensity and PR Bscans.

c. Texture features Texture features of an image quantify properties such as homogeneity, coarseness and contrast [29]. In particular, we chose to use the gray-level co-occurrence matrix (GLCM) method [28], a statistical method for analyzing texture properties of an image. It considers the occurrence of certain gray-level pixel pairs with a specific spatial relation in the image and then extracts texture features from the statistical properties of this matrix. We extracted texture features from both intensity and PR data by computing GLCM at 1-, 2-, 4- and 6-pixel distances and for the three directions of south, east and south east. For calculating each GLCM, the intensity data within the binary ROI were scaled into eight uniform levels, yielding a GLCM of size 8×8. Similarly two GLCMs for the PR image were constructed by scaling the data within the binary ROI into eight uniform levels with respect to two ranges: 1) the constant range of 0 to $\frac{\pi}{2}$ (the PR values are always bound to this range) and 2) the actual range of PR values in the

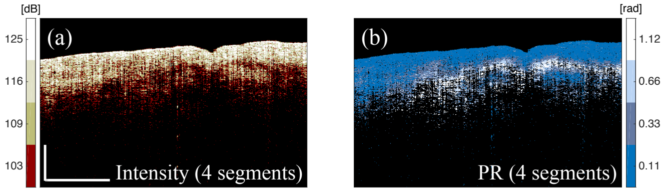

Fig. 6. Examples of four (a) intensity and (b) PR image regions calculated using *k*-means algorithm for morphological analysis. The scale bars represent 500 µm × 500 µm.

binary ROI. The following properties became the texture features [28]: 1) contrast, 2) correlation, 3) energy, and 4) homogeneity. Thus in total we calculated 48 (4 distances, 3 directions and 4 properties) and 96 (4 distances, 3 directions, 2 ranges and 4 properties) texture features from intensity and PR data, respectively.

d. Morphological features Proposed by Garcia-Allende et. al. [30] for OCT images, morphological analysis aims at statistical study of the intensity distributions of image regions. For a given Bscan, regions are constructed by segmenting the intensity values achieved by utilizing the *k*-means algorithm [31] to partition data points into *k* clusters through an iterative process. Each data point is ultimately associated with the cluster for which the data point has the smallest distance to the cluster's centroid. Figures 6(a) and 6(b) illustrate examples of intensity and PR images comprising four clusters. We extended the morphological analysis of Garcia-Allende et. al. by incorporating the positional information associated with pixels within each image region in both lateral and axial directions. The traditional and extended analysis were carried out on both intensity and PR data, because we noticed that the shape and spatial extent of the PR regions correlated with the histology-based classification of the Bscan (i.e., healthy or BCC) in some cases. For both image types, we calculated morphological features by segmenting data within the binary ROI into two through six regions. The following properties were calculated from each region: 1) mean (i.e., centroid), 2) normalized mean (i.e., centroid divided by the range of the values in the binary ROI), 3) absolute deviation (i.e., sum of absolute values of all point-to-centroid differences), 4) relative size (i.e., total number of data points in each region divided by the total number of data points in the ROI), 5) standard deviation, 6) skewness, 7) kurtosis, 8) mean of axial positions (i.e., z values), 9) median of axial positions, 10) standard deviation of axial positions, 11) mean of lateral positions (i.e., x values), 12) median of lateral positions, and 13) standard deviation of lateral positions. Thus in total we calculated 520 features (20 image regions, 13 features and 2 data types (intensity and phase retardation), based on morphological analysis.

e. Spatial frequency-based features Features computed from the 2D-DFT of an image capture the periodicity and orientation of image textures. The periodic and high PR-value bands observed in some of our PR images inspired us to extract spatial frequency-based features. The absolute 2D-DFT for each image type was first calculated [28] and then divided into regions based on spatial frequency content. We chose four regions of concentric rectangles by dividing each spatial axis into four uniform bands such that the bands have similar spatial frequencies in both axes. Feature were extracted by summing all the values in each rectangular region and normalizing them by dividing by the sum of the total value across all four regions. The features calculated from the innermost and outermost rectangular regions correspond to texture properties of the image with low and high periodicity, respectively.

2.6.3. Final set of features

Ultimately we calculated 1314 features: 613 from intensity data and 701 from PR data. The final feature matrix was sized 520 × 1314 (images x features), and the label vector describing the final classification outcome (healthy vs. BCC) had a size of 520 × 1. The feature matrix was normalized in mean and variance.

2.7. Building the classifier

Based on the features captured from the images, we constructed a classifier that predicts the presence of BCC in a set of PS-OCT images. We chose to employ supervised learning techniques, which utilize input data (i.e., *training dataset)* carrying known desired outputs (i.e., *class* or *label*) to identify the general function that maps input data to those outputs and works well for new input data (i.e., *testing data*set).

To improve the predictive power of the classifier, we followed procedures to select the best subset of features, the best machine learning algorithm, and to adjust the algorithm's parameters. Immediately, we excluded features whose values were the same for all Bscan data from all patients, since they carried no useful information for classification purpose, leaving 1144 features from 1314 original features (520 intensity and 624 PR features). To select the best subset of features, we implemented the minimal-redundancy-maximal-relevance (mRMR) algorithm [32] to identify features having high relevance (i.e., correlation) to the label vector and low correlations (i.e., redundancies) with other features. The benefits of building a classifier from fewer features are two-fold: 1) avoiding overfitting of the classifier's parameters to the current data, and 2) reducing the computational time. For each test classifier, we chose a subset of m features, $S_{mRMR,m}$, consisting of the top m high-rank features. We then tested several machine learning algorithms to determine which algorithm yielded the classifier with the highest accuracy. The algorithms tested included support vector machine (SVM) with linear and Gaussian kernels, k-nearest neighbor (KNN), and random forest. At this stage, each classifier was built using all features from $S_{mRMR,m}$ and was optimized in accuracy by adjusting its parameters. As the final step, using *forward search* algorithm [33] and the selected machine learning algorithm, we selected the final subset of features, S_{fwd}, that yielded the highest possible accuracy for the selected classifier. The forward search algorithm begins with an empty set of features (i.e., $S_{fwd} = \{\phi\}$) and upon each iteration successively adds the single feature F_i from the original feature set (i.e., $S_{mRMR,m}$) whose inclusion leads to the highest accuracy.

Instead of the standard leave-one-out-cross-validation (LOOCV) or K-fold cross-validation (K-fold CV) [34], we evaluated the accuracy of classifiers constructed at different steps using a leave-one-*patient*-out-cross-validation (LOPOCV) to avoid biasing our estimate of accuracy due to potential correlations among images from a single patient. At each iteration, the classifier is trained on feature vectors from all patients except one and tested on the remaining patient. We calculated four parameters to assess our classifier's performance: 1) accuracy, 2) sensitivity, 3) specificity, and 4) area under the receiver operating characteristic (AUC).

3. Results and discussion

We constructed three feature sets using the top 100, 200, and 300 high-rank features (sorted by mRMR): namely $S_{mRMR,100}$, $S_{mRMR,200}$, and $S_{mRMR,300}$, respectively. The numbers 100, 200, and 300 were chosen arbitrarily. For each set, we implemented various machine-learning algorithms (section 2.7) while adjusting their parameters to select the algorithm that achieved the highest classification accuracy. Ultimately, SVM with Gaussian kernel of $\sigma = 4$ (SVM_GS4) constructed from the $S_{mRMR,100}$ feature set provided the highest accuracy. For this classifier, the achieved accuracy, sensitivity and specificity were 90.2%, 91.2%, and 89.2%, respectively.

Up to this point, the selected feature subset (i.e., $S_{mRMR,100}$) had only been optimized with respect to the label vector and the original feature set, but not with respect to the selected machine-learning algorithm. Using the forward search algorithm we further optimized the feature subset by selecting only those features that were optimized for SVM_GS4 (S_{fwd}). This final future set should yield the most accurate classifier when used with SVM_GS4. We ran the forward search algorithm on $S_{mRMR,300}$ with 94 and 206 intensity and PR features, respectively.

Figure 7(a) depicts results of the forward search process as a function of the selected feature at each iteration. For all iterations, the algorithm used was SVM_GS4 and the validation method was LOPOCV. As the graphs show, the classifier's performance improves upon adding new features only until 36 features have been added (marked by the dashed black line), where the achieved accuracy, sensitivity and specificity are all 95.4%, and the AUC is 97.2%. Thus, a classifier built using those 36 features (S_{fwd}: 4 intensity and 32 PR) and SVM_GS4 achieves the best performance.

Table 3 summarizes the final 36 selected features by category. A total of 22 and 14 features were selected from Ascan-based and Bscan-based features, respectively. The 13 Ascan-based features in the short-range category and the 9 Bscan-based features in the morphological category have the highest representation. This can be attributed to either their relevance for BCC detection or to their high representation in the original feature set (Table 2). Note that exclusion of certain features in the final feature set does not in general diminish their importance or value for BCC detection: the final feature set is highly optimized with respect to the chosen machine-learning algorithm, and a different algorithm may have required a different feature set to achieve optimal accuracy. For this reason, having a large number of initial features offers the freedom and convenience of improving the classifier's performance by selecting the subset of features that is optimized for a given algorithm.

The current results are for a classifier based on both intensity and PR features. To consider whether PR data are necessary to attain high classifier performance, we built another classifier using intensity-only features and followed the same optimization steps (i.e., mRMR, algorithm selection, and forward search) to identify the most accurate intensity-only classifier. We determined a classifier built using SVM_SG5 with a subset of 18 intensity features (S_{fwd}, results are depicted in Fig. 7(b)) to yield the highest accuracy. This intensity-only classifier achieved 83.1%, 81.9%, 84.2%, and 90.8% accuracy, sensitivity, specificity, and AUC, respectively, which suggests intensity-only features yield a classifier with suboptimal performance and do not carry enough discriminatory power to detect the presence of BCC tumors. On the contrary, the strong performance of the classifier based on both intensity and PR data clearly demonstrate the power of PR data in detecting BCC tumors and thus the relevance of PS-OCT in detecting BCC in human skin.

Figure 7(c) shows the receiver operating characteristics (ROC) achieved by classifiers based on the final intensity-only or PR and intensity features. The ROCs clearly support that a classifier built based on intensity-only features underperforms the other classifier.

Table 3. Number of final selected features for each feature category.

Ascan-based features			Bscan-based features		
	Int	**PR**		**Int**	**PR**
a. Long-range	0	0	a. Basic statistics	0	0
b. Short-range	0	13	b. Histogram	0	0
c. Peaks&valleys	0	2	c. Textures	0	2
d. Segment	1	4	d. Morphological	3	9
e. Crossing	0	2	e. Spatial freq.	0	0
Subtotal	1	21	Subtotal	3	11
Total	22		Total	14	

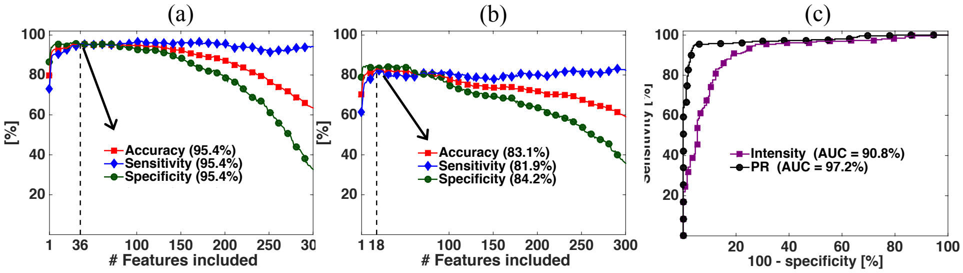

Fig. 7. Accuracy, sensitivity, and specificity calculated at each iteration of the forward search process for a classifier (a) based on both intensity and PR features, and (b) intensity-only features using SVM with Gaussian kernels of $\sigma = 4$ and $\sigma = 5$ algorithms, respectively. Vertical dashed black line represents the most accurate classifier achieved using smallest number of features. (c) ROC for classifiers built from intensity-only, and both intensity and PR features.

Although we only have demonstrated the efficacy of our proposed workflow for automated detection of BCC in human skin using PS-OCT images, we hypothesize the proposed features and the process of building and optimizing a classifier can be generalized to detecting any tissue abnormality using any type of OCT. This generalizability is mainly due to the abundant diverse features proposed and the applicability of the proposed process for building and improving a machine-learning based classifier. One disadvantage of the proposed workflow is the high computation time. Extracting features and running the forward search algorithm were the two most time-consuming steps: 17 seconds for extracting features per pair of intensity and PR images (total of 2.5 hours for all 520 PS-OCT images), and 26 hours for completing the forward search (Note that computation times are based on using a laptop with 2.5 GHz Intel Core i7 CPU and 16 GB 1600 MHz DDR3 memory). In fact, since forward search is an iterative process, selecting even the first final feature from the original feature set of 300 features takes as long as training 300 different classifiers. That said, extracting a huge number of features and running the forward search algorithm needs to be executed only during construction of the classifier; once the classifier is built, testing new samples in the clinic is quick (0.001 seconds) as they need only be tested against a small number of features (36).

In this work we did not account for the potential effect of the freezing process on the birefringence and scattering properties of our samples. Future studies may be necessary for careful evaluation of the potential changes for continued *ex vivo* work; however, we believe that the proposed workflow for automated detection is general and powerful enough to capture the presence of BCC tumors in unfrozen and even *in-vivo* human skin samples using appropriate training data.

4. Conclusions

In conclusion, we have successfully demonstrated the first automated detection of BCC in human skin using PS-OCT. Our classifier provides an unprecedented performance of 95.4% accuracy, sensitivity and specificity in detecting BCC in *ex-vivo* human skin. The proposed classifier was constructed by extracting our proposed new features in addition to features previously proposed by others from 520 PS-OCT image pairs with complementary contrasts (intensity and PR) from 42 patients.

Although our current implementation for building the classifiers is time-consuming, more efficient software platforms such as C++ or R could be used to reduce the time. Future clinical translation will require evaluation on *in-vivo* human skin samples; to accomplish this, one needs a portable and hand-held PS-OCT able to image parts of the bodies that are hard to reach by conventional stationary probes such as face, neck, and shoulder. Additionally future studies are required to focus on discriminating BCC from benign skin abnormalities (e.g., nevi), for which accessibility to a large population of diverse human skin samples is required.

Nevertheless, our results support the strong relevance of PS-OCT for detecting BCC in human skin, as our proposed classifier outperformed even the best possible classifier based on intensity-only images. Additionally our proposed workflow for automated detection can be generalized to detecting any other tissue abnormalities using any type of OCT system.

Funding

This work was partially funded by a Stanford Bio-X Seed Grant (IIP6-87). Tahereh Marvdashti was partially financially supported by a Stanford Graduate Fellowship.

Acknowledgments

We would like to thank Owen Ma, Jamaica Jensen, Samantha Rodriguez, and Katherine Ransohoff of the Stanford dermatology clinic for their help in collecting human samples.